Procedia Computer Science

Volume 51, 2015, Pages 1705–1713

ICCS 2015 International Conference On Computational Science

Steps Towards Bridging the HPC and Computational Science Talent Gap Based on Ontology Engineering Methods

Svetlana Chuprina
Perm State University, Bukireva Str. 15, 614990, Perm, Russia
chuprinas@inbox.ru

Abstract
The paper describes an ontology-based methods and framework for design of learning courses covering the HPC and Big Data areas and how to include these into Computational Science training within the remit of existing courses of Master Programme entitled "Applied Mathematics and Computer Science" (Faculty of Mechanics and Mathematics, Perm State University, Russia). It helped bringing together the university and IT-companies around a real industry projects in the field of Big Data with active participation of master's students. In this paper, the visual tools and ontology-based methods for computer-supported collaborative learning environment will be also presented.

Keywords: Big Data, HPC, talent gap, ontologies, ontology engineering methods, collaborative learning, contextual semantic search, adaptable visual ontological editor

1 Introduction

Nowadays Big Data analytics is an area of rapidly growing technological diversity. Changes in the curricula of relevant areas of training in universities do not keep pace with the rapid growth in demand for specialists in the area of Big Data and rapid hardware and software platforms' updates in HPC. The success of Big Data initiatives will depend upon the ability of companies and organizations to find the right talent, as well as their ability to effectively integrate this talent into the business infrastructure. According to the often-quoted definition of Big Data we define Big Data technologies as a new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis (IDC, 2012).

To help tackling the problems mentioned above we must try to find a comprehensive solution based on unique conceptual platform, which is adequate to students' backgrounds from different

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2015

doi:10.1016/j.procs.2015.05.308

faculties. It is a very complex problem because educational process in the universities involves a big amount of very different departments, which is also present in our university. The structure of Perm State University (PSU) comprises 12 faculties (Biology, Chemistry, Economics, Geography, Geology, History and Politology, Law, Mechanics and Mathematics, Modern Foreign Languages and Literatures, Philosophy and Sociology, Physics) and every faculty has a number of departments and several programmes for Bachelor and Master Degree. Totally there are 80 departments in PSU which are also responsible for the postgraduate studies and the research work of the staff in very different research areas. How to find in this case the common conceptual platform in context of interdisciplinary projects in Big Data area with students' participation?

Working with big data involves proficiency in number of disciplines. At least, big data analytics requires skills in mathematics, data analysis, statistics, data cleaning, applied problem solving combined with programming and computational skills. Many universities are launching new programmes in Big Data and Data Science in order to meet the growing demand. In particular, there are several educational programs in Mechanics and Mathematics faculty of PSU directly aimed at training students in the field of HPC. But to achieve success in interdisciplinary projects it is necessary to change the existing education programs in a big variety of different areas of training in university. We believe that, one way or another, this must affect the educational process at all faculties of the university, and not just those who teach students in IT areas only.

This paper presents how the problems mentioned above have been tackled within the remit of existing courses of Master Programme entitled "Applied Mathematics and Computer Science" in Perm State University. It demonstrates the results of ontology designing within Adaptable Visual Ontology Editor ONTOLIS environment and the usage of designed ontologies in student learning projects as well as in industry projects.

2 Bridging the HPC and Computational Science Talent Gap

The concept of integration of the basic part of different undergraduate programs in Computer Science and Information Technologies in the universities conducting training in several IT-programs is suggested. Analysis of relevant educational programs showed that at the initial stage of training this integrated base part may provide the formation of general and professional competencies of the students of these categories of up to 40% for bachelor's programs. The concept is realized in the Perm State University for set of six educational programs on bachelor in computer science and information technology (Rusakov S., Khenner E., & Chuprina S., 2014).

Because Big Data, inter alia, is the process of changing data into information, which then changes into knowledge, and every learning discipline can have her own application domain, the development of student skills in ontology design can be considered as a common conceptual platform for collaboration within interdisciplinary projects and a common repository of application domain ontologies can be developed by students from different departments. The ontology design process should be managed under a highly qualified supervision. The supervisor may be from university, as well as from industry.

To use the common repository of application domain ontologies as a platform for collaboration is reasonable because one of the more common goals in ontologies developing is to specify in explicit form common understanding of the structure of information from some application domain among people or software (Gruber T. R., 1993). Therefore, we offer all the faculties to introduce the special issues into existing training courses in computer science and/or database area devoted to designing and development of application domain ontologies related to fields of different faculties' research and education interests. It will be a good conceptual basis to support the participating students from different faculties in interdisciplinary projects and to help to provide the collaboration between

academia and industry around real-world projects. It is a basis for the integration of different existing learning tools and systems in a common learning environment in the future.

The computer science department of Mechanics and Mathematics Faculty plays a special role in this process. A big volume of teaching load of this faculty staff has already included the computer science teaching in various faculties and departments. Because of the diversity in students' backgrounds, each syllabus must be tailored to the student's needs based on experience and interest, as well as available faculty and courses. To simplify the ontology developing process in adaptable way in accordance with the personal preferences of the students with different backgrounds we have developed an adaptable visual ontology editor.

To pair Ontology Engineering with Big Data techniques and technologies within educational process in Computer Science Department we have done a lot of changes in our own curricula and syllabuses. To make the changes in curricula firstly we have reviewed the local industry proposals for the introduction of new courses and programmes in Big Data and HPC study areas and after that we have developed the amendments to the existing courses and programmes and made some restrictions of existing courses and programmes and closed others.

The Bachelor Programme "Applied Mathematics and Computer Science" additionally to fundamental mathematical disciplines includes "Parallel architectures", "Parallel programming", "Distributed algorithms" and the Master Programme (of the same title) includes "HPC and Grid technologies". In order to suite the local industry proposals we have to focus on the topics related to Big Data problems.

Besides the topics with common definitions and main conceptual elements among the proposals for the introduction of new Big Data technologies courses there are the following:

- Tackling Big Data problems with HPC;
- MapReduce paradigm and programming model;
- Apache Hadoop platform;
- Apache Hive data warehousing component;
- Apache Spark and Spark SQL;
- Machine Learning with Apache Mahout;
- Big Data technologies as cloud-based solutions;
- Capabilities of NoSQL systems.

To use Big Data technologies for analytics it is needed to master at least data mining, multi-dimensional analysis and data visualization. Some applications of Big Data technologies are used not only for analysis of data, but also for operations (deploying Web sites for social media, gaming applications, processing online orders and so on) and for large content stores that provide information access to massive amounts of documents. To meet the curricula time limits we were not able to introduce the new course covering all these topics, therefore we have developed a corresponding amendments to the syllabus of following existing courses on the Master Programme:

- Intelligent Systems;
- Advanced Internet Technologies;
- Knowledge Discovery and Data Mining;
- Research Seminar.

To cover the other important aspects of HPC, two years ago we introduced the Master Programme "Applied Mathematics and Computer Science" the new course titled "Advanced algorithms".

Ontologies developed under master's educational process (for example, within Intelligent Systems course) are used not only as means to better study and systematization of knowledge in Big Data areas but also as a basis for better understanding among the participants of joint interdisciplinary projects, as well as artifacts which may be used as a part of content of ontologies repositories suitable to reusing within Big Data projects. As an example of such a project, we can refer to the successful participation

of our students in the project ONTOC2R. Project ONTOC2R was devoted to the development of a demonstration prototype text mining platform with the possibility of semantically meaningful search in potentially unlimited amounts of unstructured information based on domain ontologies and cloud platform repository C2R (Kostarev A., 2014).

The research results of some of our students, achieved in the frame of ONTOC2R project, are reflected in the following publications (Pleshkova I., 2015) (Postanogov I., 2015).

Now three of our MSc students take part in research project entitled "Development Models and Tools to Transform Traditional Information Systems to an Intelligent Systems via use of a Bilingual Ontology in the Computing Area". The reported study was partially supported by the Government of Perm Krai, research project No. C-261004.08.

This project was developed with the participation of the author of this paper. Despite of the significant progress in the domain of modern intelligent IT, widely available unified solutions for semantic search, adequately aware of the context of the retrieval request, are still missing. Featuring such unified mechanisms into developed and existing information systems will improve the quality of solving a wide variety of problems related to pertinent information retrieval. Among such problems, there are filtering of unwanted mailing, automation of analytic reports construction, automatic generation of teaching information materials, specific domain text synthesis, etc.

Such instruments will allow transforming the traditional information systems based on relational model into intellectual ones based on the methods and techniques of ontological engineering and external linguistic resources reusing. An outstanding feature of the suggested approach within the mentioned above project is the possibility of automatic handling of the extensible ontology of specific domain extracting from the database of legacy information system without source code modification. The enriching of semantic search services is possible due to taking into account synonyms, hyperoniums and hyponyms in the context of the retrieval request.

The proposed unified solution will be demonstrated by the developing of extensible bilingual knowledge base in the domain of Computing to increase the semantic power of search engines.

As international users increase rapidly, multilingual systems have become a very important service for global users. Unlike thesauruses and electronic bilingual definition dictionaries, which translate and explain concepts using the context predetermined by their authors, the proposed approach allows dynamically creating new descriptions and automatically selecting required context of the explanation immediately in the process of the retrieval and automated information analysis. For these purposes we propose to use the modern approaches from Data and Text Mining. One of the essential results will be designing the model and creating the bilingual (Russian and English) ontology containing the basic concepts in the domain of Computing in OWL standard. This will allow not only export/import and reusing of the ontology within the existing and developed in the future information systems, but improving the quality of text translation, because nowadays there is lack of available instruments for adequate context-aware matching between English and Russian concepts in the domain of Computing.

So, we present the approach not only for development of new intelligent information systems but for the automated transformation of existing traditional information systems to intelligent, without requiring changes to the source code of the legacy information systems. This transformation is necessary in order to reuse knowledge extracted from databases and heterogeneous e-learning resources for semantic retrieval and user activity automation. The proposed approach based on ontology, topic maps and metadata representation standards shall improve the interoperability and adoptability of the designed software tools.

The main feature of ontologies used within the learning process is that it is not only the subject of study, but also the artifact ready to be a basis for research tools development, the mean of collaboration. If the ontology is represented in standard format it enables semantic search and reasoning using standard, freely available third-party tools. The most complicated stage is the building of application domain ontology, however once it is created, it can be reused for solving a wide variety

of this application domain problems. For that reason we collected all application domain ontologies developed by students in common repository.

3 Visualization Tools and Adaptable Visual Ontology Editor ONTOLIS

The visualization tools are the essential part of collaborative learning. We use not only third-party software for these purposes, but also the systems developed by our graduate and post-graduate students within real-world projects. These tools visualize both the working process and the results within project-based and collaborative learning. We use the open source software Gource (Gource, 2014) to monitor the activity of the students according to the history obtained, for example, from version control system that tracks their projects. Figure 1 shows the activities of the students while they are solving different tasks on one collaborative project.

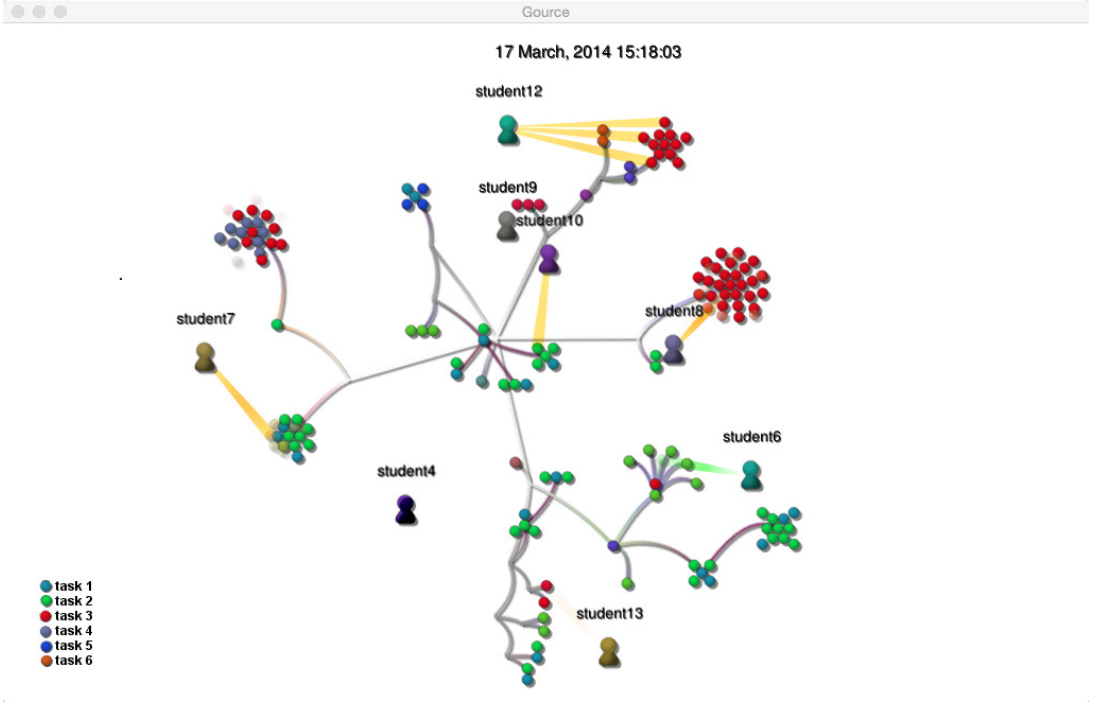

Figure 1: Students' activity visualization within Gource

The history mentioned above should be converted into the Gource log format (based on XML). There are standard converters for popular version control systems (for example, git, svn, etc.), but in the case of other history journal the user will have to implement the converter by himself. Besides Gource is crossplatform but not multiplatform application: it can run under Windows, GNU / Linux и OS X, but there is no version for mobile devices. That is why we research another way to visualize the projects' activities using our own multiplatform visualization software system SciVi (Ryabinin K. & Chuprina S., Adaptive Scientific Visualization System for Desktop Computers and Mobile Devices, 2013) (Ryabinin K. & Chuprina S., Development of Multiplatform Adaptive Rendering Tools to Visualize Scientific Experiments, 2014) that has special tools to adapt to different solvers and sets of

data. In this case the solver is the version control system that produces the necessary log file with the history of students' activities.

As for the visualization of projects' results, we usually use visual tools to tackle ontology engineering problems. According to the Master Programme in our department it is actually important to use the artificial intelligent means in the projects that are developed by the students. This corresponds to the needs of the modern IT industry, therefore we have collaborative projects with IT-companies. The students have opportunity to learn and work at the same time. The main feature of some of these projects is using ontology engineering methods, data and text mining tools to develop context-aware search engines adaptive to the wide sphere of different application domains.

MSc students from Computer Science Department in collaboration with their teachers have developed the adaptive ontological visual editor called ONTOLIS (Chuprina S.I. & Zinenko D.V., 2013) to simplify the ontology developing process in adaptable way and in accordance with the personal preferences of students with different backgrounds. Moreover, we used this tool to improve the adaptive capabilities of SciVi. The adaptation module of SciVi uses different ontologies, for example the ontology of visual objects. This ontology is used to adapt SciVi to different visualization techniques. The fragment of this ontology within Ontolis environment is shown in Figure 2.

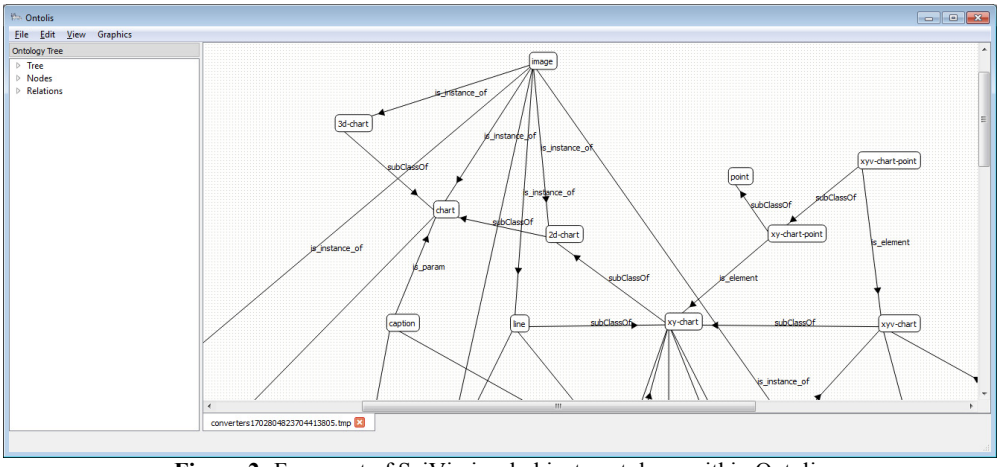

Figure 2: Fragment of SciVi visual objects ontology within Ontolis

ONTOLIS is a multi-platform software tool for visual creating and editing of ontologies in adaptable way. This editor implements the original algorithm to adapt the layout capabilities and visual settings format of the visual editor to the specifics of the domain ontology. It is allowed to customize the visual editing environment to processing of ontologies presented in standard format (OWL, RDF) as well as in the custom format. ONTOLIS is adaptable to the representation of an extensible set of different types of entities visualization features (a certain type of shape, color, etc.) and relationships between the entities by means of metadata. The business logic associated with the interpretation of the ontological knowledge is completely separated from the ontology editor. OWL export / import is available.

The same ontology fragment within popular ontology editor Protégé is shown in Figure 3. However Ontolis is more suitable for our needs thanks to its high-level adaptation means based on meta-ontology.

Bellow Figure 4 and Figure 5 present the results of Data Base Models ontology (fragment in Russian) and Big Data ontology (fragment in Russian) development with different ONTOLIS settings according to the students' personalized preferences.

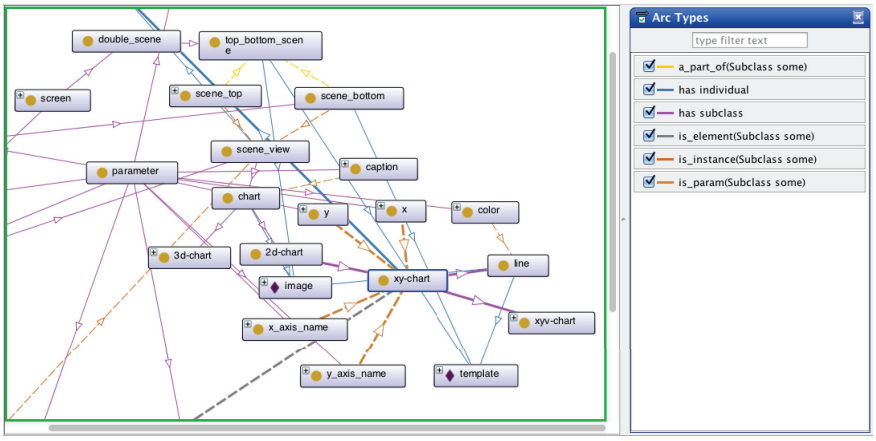

Figure 3: Fragment of SciVi visual objects ontology within Protégé

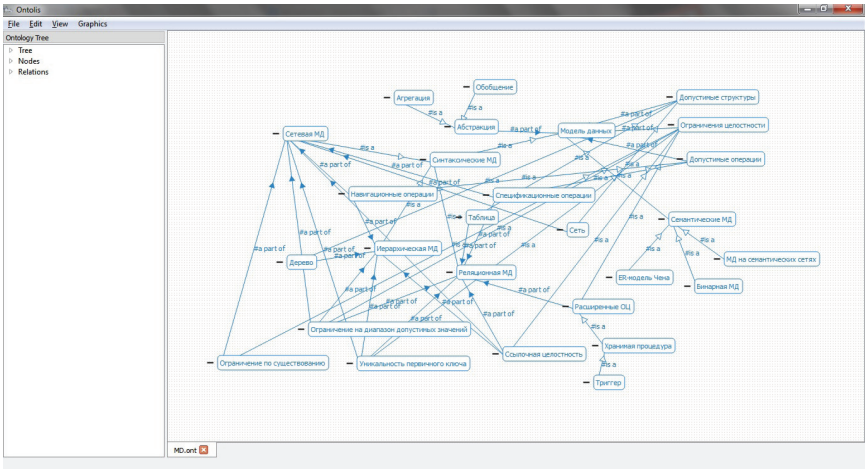

Figure 4: Fragment of students Data Base Models ontology within ONTOLIS

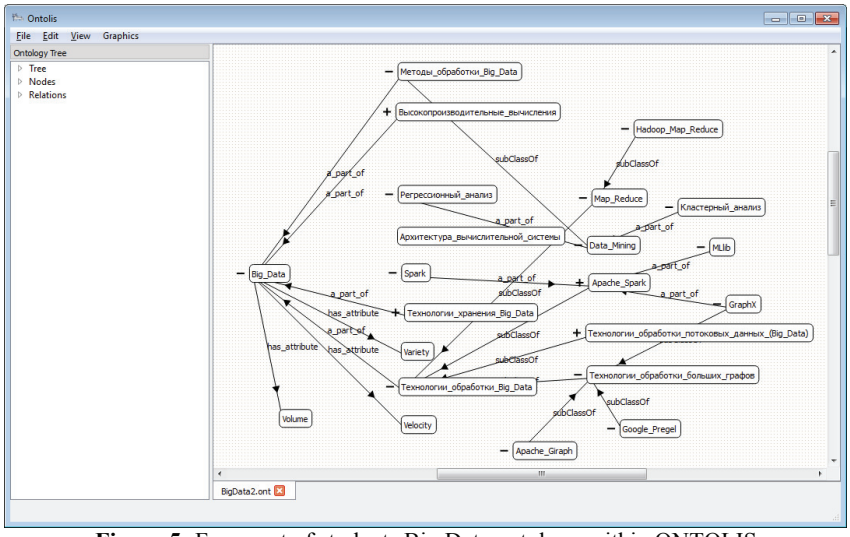

Figure 5: Fragment of students Big Data ontology within ONTOLIS

4 Future Work

We hope that the ontologies repository developed by MSc students will be an important part of future intelligent Learning Management System (LMS), because nowadays we have only traditional LMS in PSU called ETIS and there are no capabilities to do semantic contextual search within the educational documents store (Chuprina S. & Statsenko N., 2010). We are planning to build such intellectual LMS using ETIS as a legacy information system based on the principles that have been used for designing of DEPTHS (Design Patterns Teaching Help System) environment. DEPTHS environment is an open-source project. In the paper (Jeremic Z., Jovanovic J., & Gasevic D., 2009), it is presented how to develop and apply a common ontological foundation for the integration of different existing learning tools and systems in a common learning environment. In DEPTHS, Learning Object Context Ontology (LOCO) framework is proposed as an ontology base for the integration. This is a generic framework capable of formally representing diverse learning contexts. Accordingly, the framework integrates a number of learning-related ontologies, such as a user model ontology, a learning content ontology, and domain ontologies.

We think of Big Data as a big opportunity to develop the next generation of technologies to store, manage, analyze, share, and understand the huge quantities of data we are now collecting. Based on interactions with our industry partners, from the Small Innovative Enterprise (SIE) KNOVA, we have gained a unique perspective on the issues posed by large amounts of complex, digital data. Through the tackling of the challenges of Big Data, participants of SIE will have the opportunity to learn about the most recent research developments from the scientists behind the work, and gain key insights into the important new developments that will shape the landscape of data processing in the years to come.

5 Conclusion

In this paper, we have tried to show how to help to narrow the gap between traditional students IT skills and recently demanded analytics and storage skills in the world of big data via syllabus and curricula changes in the higher educational setting.

The adaptive visual ontology editor ONTOLIS is presented as a platform that enabled putting together a local IT business company and master students from Perm State University within Small Innovative Enterprise. The collaborative work and discussions on the opportunities and challenges in ontology designing helped to design IT real-world projects in Big Data processing area with master students participation to develop the context-aware semantic search tools in huge structured and unstructured data sets. Students' participation in these projects and the related results as well as the grant research funding received by students have already proven the validity of our concept how to bridge the HPC and Computational Science talent gap based on ontology engineering methods.

References

Chuprina S., & Statsenko N. (2010, November 4-5). Using Ontology and Metadata to Integrate eLearning Resources and Administrative Information System of University. *Proceedings of the 9th European Conference on E-Learning* , 171-179.

Chuprina S.I., & Zinenko D.V. (2013). Adaptable Visual Ontology Editor ONTOLIS (in Russian). *Vestnik of Perm State University , 3* (22), 106-110.

Gource. (2014, October 16). Retrieved February 21, 2015, from code.google.com/p/gource/: https://code.google.com/p/gource/

Gruber T. R. (1993). A translation approach to portable ontology specification. *Knowledge Acquisition , 5* (2), 199-220.

IDC. (2012, March). *Worldwide Big Data Technology and Services 2012-2015 Forecast, IDC #233485*. Retrieved February 21, 2015, from www.idc.com: http://www.idc.com/research/viewtoc.jsp?containerId=233485

Jeremic Z., Jovanovic J., & Gasevic D. (2009). Evaluating an Intelligent Tutoring System for Design Patterns: the DEPTHS Experience. *Educational Technology & Society , 12* (2), 111–130.

Kostarev A. (2014, January 25-26). Cloud Content Repository C2R: Text Analysis Based on Ontology Engineering Methods (in Russian). *Proceedings of the 9th Conference "Free Software in High School"* , 23-28.

Pleshkova I. (2015, January 24-25). Using the Genetic Algorithm to Enhance Semantic Search Quality Within a Large Amount of Unstructured Documents to Improve Scientific Research Results (in Russian). *Proceedings of the 10th Conference "Free Software in High School"* , 50-52.

Postanogov I. (2015, January 24-25). Using Apache Spark Technology Stack and Ontology Engineering Methods to Develop Intelligent Analysis of Complex Network Big Data (in Russian). *Proceedings of the 10th Conference "Free Software in High School"* , 52-57.

Rusakov S., Khenner E., & Chuprina S. (2014). Integration of Basic University Training of Specialists in Computer Science and Information Technologies (in Russian). *University Management: Practice and Analysis* (3), 119-125.

Ryabinin K., & Chuprina S. (2013). Adaptive Scientific Visualization System for Desktop Computers and Mobile Devices. (Elsevier, Ed.) *Procedia Computer Science , 18*, 722-731.

Ryabinin K., & Chuprina S. (2014). Development of Multiplatform Adaptive Rendering Tools to Visualize Scientific Experiments. *Procedia Computer Science , 29*, 1825-1834.